

FinUgRevita: Developing Language Technology Tools for Udmurt and Mansi

Veronika Vincze¹, Ágoston Nagy², Csilla Horváth², Norbert Szilágyi³,
István Kozmács³, Edit Bogár², Anna Fenyvesi²

¹University of Szeged, Department of Informatics

²University of Szeged, Institute of English-American Studies

³University of Szeged, Department of Finno-Ugric Studies

finugrevita@gmail.com

December 16, 2014

Abstract

Nowadays, digital language use such as reading and writing e-mails, chats, messages, weblogs and comments on websites and social media platforms such as Facebook and Twitter has increased the amount of written language production for most of the users. Thus, it is primarily important for speakers of minority languages to have the possibility of using their own languages in the digital world too. The FinUgRevita project aims at providing computational language tools for endangered indigenous Finno-Ugric languages in Russia, assisting the speakers of these languages in using the indigenous languages in the digital space. Currently, we are working on two Finno-Ugric minority languages, namely, Udmurt and Mansi. In the project, we have been developing electronic dictionaries for both languages, besides, we have been creating corpora with a substantial number of texts collected, among other sources like literature, newspaper articles and social media. We have been also implementing morphological analyzers for both languages, exploiting the lexical entries of our dictionaries. We believe that the results achieved by the FinUgRevita project will contribute to the revitalization of Udmurt and Mansi and the tools to be developed will help these languages establish their existence in the digital space as well.

This work is licensed under a Creative Commons Attribution–NoDerivatives 4.0 International Licence.
Licence details: <http://creativecommons.org/licenses/by-nd/4.0/>

1 Introduction

In the age of modern technology, the constant development and widespread usage of technical tools such as the internet and smartphones enable people to communicate in real time throughout the world. Human-human interaction and machine-human interaction is supported by several language technology tools and applications such as spellcheckers, machine translation websites and search engines, besides, online resources and databases are exploited in communication in the digital world. However, the fact that while there are effective language technology tools available for languages with millions or billions of speakers, for minority languages even the most basic digital language processing tools are often missing. Hence, it is of utmost importance to develop language technology tools for users of minority languages, in order to facilitate communication in their mother tongue in the digital world as well.

Minority languages differ from other languages not only with respect to the numbers of their speakers but with respect to the fact that they are usually not recognized as official languages in their respective countries, where there is an official language and one or more minority languages. Thus, it is often the case that the speakers of minority languages are bilingual, and usually use the official or majority language at school and at work, and the language of administration is also the majority language. On the other hand, the use of the minority language is typically restricted to the private sphere, i.e. among family and friends, and thus it is mostly used in oral communication, with only rare examples of writing in the minority language.

Nowadays, digital language use such as reading and writing e-mails, chats, messages, weblogs and comments on websites and social media platforms such as Facebook and Twitter has increased the amount of written language production for most of the users [1]. Thus, it is of primary importance for bilingual speakers to be able to use their mother tongues in the digital space as well (cf. [2]).

In order to implement user-friendly language technology applications such as the above-mentioned spellcheckers or machine translation systems, basic linguistic pre-processing technologies are a must for the given language. In the case of minority languages, natural language processing might encounter problems even at the level of character encoding, provided that there are no standardized or well-known character sets in use. For higher-level language technology applications, it is further necessary to have a sentence splitter and tokenizer, a morphological analyzer and part-of-speech tagger, moreover, to get a deeper understanding of the content of texts, syntactic and semantic parsers are indispensable. These tools are often used in a chain: for instance, the output of the tokenizer is the input of the morphological analyzer, and the syntactic parser usually makes use of the output of the POS-tagger when parsing sentences.

In this paper, we discuss work within our project, FinUgRevita, which seeks to

create language technology tools for minority Finno-Ugric languages. We first describe the project, then we provide some basic background to the languages we are currently working on: Udmurt and Mansi. Later, we present the main tasks of the project, i.e. corpus building, developing electronic dictionaries and morphological analyzers. Lastly, we offer some possible directions for future work that we intend to do in the next phases of the project.

2 The FinUgRevita Project

The FinUgRevita project¹ aims at providing computational language tools for endangered indigenous Finno-Ugric languages in Russia, assisting the speakers of these languages in using the indigenous languages in the digital space, and assessing, with the tools of sociolinguistics, the success of these computational language tools. The project is supported by the Hungarian National Research Fund and the Finnish Academy of Sciences, and is carried out by researchers working at the University of Szeged and the University of Helsinki.

In the computational linguistic component of this project we plan to use existing language resources in endangered minority Finno-Ugric languages to develop computational tools (learning tools and authoring tools) that would enable speakers to use their minority language in modernized popular discourse required in common everyday functions of written language use. Another key goal of the project is to provide these tools free of charge to anyone who is interested in learning and practising these languages. The tools, we believe, will increase speakers' proficiency in their minority language, positively change speakers' attitudes to their minority language, and, in the end, aid the revitalization process.

3 The Languages: Udmurt and Mansi

Here we provide some background on Udmurt and Mansi and basic demographic data on their speakers.

3.1 Udmurt

The Udmurt language (or, by an earlier exonym, Votyak) is a member of the Uralic language family, a somewhat endangered indigenous language in Russia. It is spoken in the area between the Vyatka, Cheptsa and Kama rivers, about 1,200 kilometers

¹<http://www.ieas-szeged.hu/finugrevita/index.html>

(about 750 miles) east of Moscow but west of the Ural mountains, in the Udmurt Republic (or, informally, Udmurtia). Additionally, Udmurts also live in greater numbers in Kazakhstan, and dispersed in many cities and towns of Russia. According to the latest, 2010, Russian census, 552,299 people profess to be of Udmurt ethnicity and 324,338 to be speakers of the Udmurt language. (Both figures have been decreasing from census to census in recent decades.)

Today, the Udmurt language is used mostly within the family and among friends, and even though it is an official language in Udmurtia, it has limited power and rights. It is not used in the legislature or political life. However, it is present in the media, education, and the cultural sphere, as well as enjoying a growing presence on the internet.

3.2 Mansi

The Mansi language (or, by an earlier exonym, Vogul) is a member of the Uralic language family, a severely endangered indigenous language in Russia. It is spoken primarily in the Khanti-Mansi Autonomous Okrug of Western Siberia. According to the latest, 2010, Russian census, 12,269 people profess to be of Mansi ethnicity and 938 to be speakers of the Mansi language. (The former figure has been increasing from census to census in recent decades, while the latter decreasing.)

Today, the Mansi language is used mostly within the family and among friends. It has no official status or economic value associated with it. It is not used in the legislature or political life. However, it is present in the media, education, and the cultural sphere, as well as enjoying a growing presence on the internet.

4 A Survey of User Data: The Case of Saami

At the beginning of our project, we contacted the maintainers of the website Giellatekno², which offers many important CL resources and tools for several minority languages including various dialects of Saami, Circumpolar and Uralic languages. They kindly provided us their access logs, on the basis of which we were able to carry out some quantitative data analysis in order to gain some insight into what user preferences are when using CL resources and tools for minority languages.

First, we analyzed dictionary searches made in Giellatekno's database. It was revealed that the most frequently searched language pairs are Northern Saami – Norwegian and vice versa, Northern Saami – Finnish and vice versa, Finnish Kven – Norwegian, Nenets – Finnish and Western Mari – Finnish. The users usually seek to

²<http://giellatekno.uit.no/>

translate words from Northern or Southern Saami, Finnish Kven or Nenets, on the other hand, the languages they would like to translate into are usually Norwegian, Finnish or English. All this suggests that most users translate from a minority language to a majority language (or a widely known second language like English), with the exception of Saami dialects, where both translation directions are widely attested. The number of page visits also demonstrates that online dictionaries play an essential role in learning minority languages. With this in mind, we felt it necessary to set ourselves the goal of creating online dictionaries for both languages we are working with (see Section 5.1 for details).

Second, we also analyzed the demographic data of the users of the page. We were also given access to the Google Analytics of the Giellatekno sites. Most of the users of the GT site still use Norwegian (Bokmål) on their computers. In the last month (Oct 2014), 10,000 people connected to the site, and more than 6,000 of them use Bokmål, while the second most important language is English with 1,300 users, and the third is Finnish with 1,000 users.

Google Analytics also provide data about the location of the access. These are in line with the language data: most of the users connect to the site from Norway (8,000), the second one is Finland with 1,400 users and the third is Sweden with nearly 600 users. All this proves that existing online resources for Finno-Ugric languages raise the interest of users across linguistic and geographic boundaries, which tendency we would also like to exploit in our project, that is, we intend to make our resources freely available on the web.

5 FinUgRevita's Contributions

In this section, we present the FinUgRevita project's most important contributions to the computational linguistic field, which cover the digitization of existing resources and the implementation of new tools and resources as well.

5.1 Creating online dictionaries

The creation of online electronic dictionaries is in progress for the two main languages of the project, Mansi and Udmurt.

The original paper-based Udmurt–Hungarian dictionary we are using as a starting point was compiled and edited by István Kozmács ([3]). In the project, the electronic version (Microsoft Word document) of this book is used and is transformed for our needs semi-automatically. First, the document is transformed into a simplified HTML containing the main text style character markers (like **bold** or *italics*). On the basis of

this formatting, the whole document is converted into a CSV file (comma-separated values) automatically, but this has to be reviewed manually since a paper-based dictionary contains some shortcuts which do not enable its automatic processing, for instance, it contains coordinations that can only be interpreted by humans. At this stage, the automatic conversion has been already carried out, and the manual correction phase is in progress. The dictionary contains approximately 13,000 entries.

The project's online Mansi dictionary is going to be based primarily on the already existing Mansi–Russian and Russian–Mansi dictionaries, compiled by Mansi scholars. The online dictionary covers the lexical material of Rombandeeva's and Kuzakova's dictionary [4], and Rombandeeva's Russian–Mansi dictionary [5], collated with the data of Munkácsi's enormous Mansi–Hungarian dictionary [6] and also expanded with the Northern Mansi material of Balandin's and Vakhrusheva's Mansi–Russian dictionary [7], as well as with dozens of the most necessary neologisms describing different features of contemporary lifestyle (such as the urban environment, oil mining or judicial terms), created and used first and foremost by the journalists of the Mansi newspaper *Luima Seripos*.

The beta version of the online Mansi dictionary will contain approximately 10,000 entries. The Mansi lexemes will be supplemented with English, Russian and Hungarian translations, parts of speech and annotation of the sources, i.e. the dictionaries that are contained within. The Mansi forms are retrieved from the PDF versions of the dictionaries by means of optical character recognition, while the English and Hungarian translations are provided by linguists. Figure 1 presents the process of dictionary building: the automatic optical character recognition is followed by manual correction and translation of the entries, and then this database is turned into a searchable, digitized dictionary [8].

The online Mansi dictionary being a key resource for creating a morphological analyzer, the project also aims to make it available for public use as well, thus meeting a long-felt need for a sufficient Mansi–English–Mansi and a suitable online Mansi dictionary.

5.2 The Development of Morphological Analyzers

One of the most important tasks of this project is to create morphological analyzers. First, morphological analyzers for the Finno-Ugric languages we are working on were searched for and their usability was evaluated.

For Mansi, we were able to find a morphological analyzer [9] developed by MorphoLogic Ltd.³. However, it was not applicable to our purposes for several reasons.

³http://www.morphologic.hu/urali/index.php?lang=hungarian&a_lang=chv

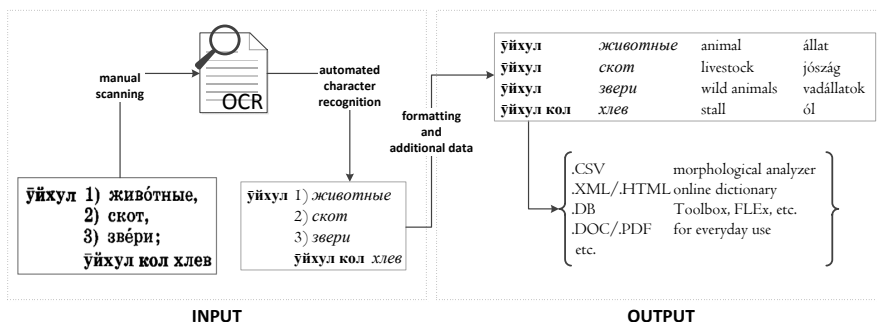


Figure 1: The process of dictionary building

First, it employs Latin-based transcription but the current Mansi orthography is Cyrillic-based (see Section 5.3). Second, its vocabulary completely lacks the contemporary lexicon of the 20th and 21st centuries since it is based on Munkácsi’s Mansi dictionary [6] and it was optimized for the texts covered in Kálmán’s *Chrestomathia Vogulica* [10] and *Wogulische Texte* [11], mostly collected at the end of the 19th century. Third, it is not open-source. For all these reasons, we decided to create a new morphological analyzer for Mansi from scratch. The dictionary mentioned in Section 5.1 will serve as a basis for the morphological analyzer as well, and lexical entries of Mansi are now being grouped into different morphological categories depending on the conjugational/inflectional paradigm they belong to. For this, we rely on the descriptions found in several Mansi grammars [12, 13], as well as on the linguistic intuitions of native speakers of Mansi.

In the case of Udmurt, we contacted the developers of the already existing Udmurt analyzer available at <http://giellatekno.uit.no/cgi/d-udm.eng.html>. We collaborate now with them and our task is mainly to correct and to create the lexical database and the grammatical rules behind the analyzer. The lexical material

| Text type | Number of characters | Number of words |
|------------|----------------------|-----------------|
| Blogs | 26,615 | 3,969 |
| Wiki | 32,110 | 4,293 |
| Literature | 142,272 | 20,899 |
| Newspapers | 216,740 | 30,664 |
| Education | 49,294 | 6,897 |
| Essays | 25,388 | 3,255 |

Table 1: Proportion of text types in the Udmurt corpus

of our Udmurt dictionary mentioned in Section 5.1 is also being integrated into the database of the morphological analyzer.

5.3 Corpus Building

In order to create and test the applications to be made in the project, corpora of Mansi and Udmurt are being created. The corpora contain mainly newspaper articles and literature, but other types of texts are also planned to be integrated. Now, raw texts are collected, and later these texts will be transformed into a uniform structure and annotated.

Table 1 summarizes the number of words and characters in each discourse type of the Udmurt corpus. As can be seen, the biggest represented text type is the newspaper section with the published available volumes of the Udmurt language periodical *Udmurt Dunne*, but material from some children’s journals like *Kizili* and *Zechbur* and other newspapers are also included here. Topics vary from interviews to sports and cultural news, reports on events etc.

We were also able to collect material from the web, i.e. Wikipedia pages and weblogs, due to the growing presence of the Udmurt language in the social media as well. We also included some academic essays in the corpus, together with texts on education. Most of these texts were already digitized, which made it easier for us to collect and process them. The corpus now contains approximately 70,000 tokens.

The core of the Mansi corpus consists of the articles published in the Mansi newspaper *Luima Seripos*. The editorial staff of *Luima Seripos* (Mansi for “Northern dawn”) separated from the regional minority newspaper and started the Mansi monolingual newspaper on 11 February 1989. The length of the newspaper started from two pages, appearing twice a month, then increased to eight pages per week, and it has recently been published on sixteen pages every two weeks. The online archive of *Luima Seripos*, consisting of 46 issues, is available on the homepage of the joint editorial board

of *Luima Seripos* and regional Khanty newspaper *Khanty Yasang*.⁴ This database, together with several former issues, increases the project's Mansi corpus up to 260 exemplars, that is, to approximately 5,200 articles. The corpus now contains more than 1 million tokens.

The Mansi texts published in *Luima Seripos* cover various topics, most importantly not only those introducing traditional lifestyle, folklore and short biographies, but domains of urban life as well, thus they provide the project with a multilayered and diverse corpus. Since the Mansi newspaper is the only stable and complex source of Mansi texts, of all the possible sources it has the greatest impact on the language use of the Mansi population.

Using *Luima Seripos* as the primary source of the Mansi corpus also defines the project's choice for Mansi orthography. The first researchers visiting the Mansi used different Latin-based transcriptions to write down Mansi texts, and the first attempts to create the standard variety and orthography for the Mansi language at the beginning of the Soviet era were based the Latin alphabet as well. Cyrillic transcription came into use in 1937 when all the nationalities living in the Soviet Union were ordered to switch over to the use of Cyrillic-based alphabets. The change caused several problems and the unsuitability of the Cyrillic alphabet and orthographical system to represent the morpho-phonological features of the Mansi language was not the smallest among them. The newspapers, schoolbooks and other works published in Mansi were inconsistent in marking special phonemes (such as the grapheme *ɣ* denoting the phoneme *ŋ*), or vowel length (despite of its role in differentiating the meaning of words, e.g. *oc* 'surface' and *ōc* 'sheep'). Nowadays the Mansi writing system is almost completely unified [14], the only minor difference between the two currently used orthographies is marking the palatal fricative: while scientific works use a combination of letters *c* and palatalizing vowels, in non-scientific publications, such as the *Luima Seripos* newspaper, and, for instance, schoolbooks in alternative educational institutions the authors replace *c* with *ɰ*.

6 Summary

In this paper, we have discussed the FinUgRevita project, which seeks to provide language technology tools for two Finno-Ugric minority languages, namely, Udmurt and Mansi. Currently, we have been developing electronic dictionaries for both languages, besides, we have been creating corpora with a substantial number of texts collected, among other sources like literature, newspaper articles and social media. We have

⁴<http://www.khanty-yasang.ru/luima-seripos/archive>

been also implementing morphological analyzers for both languages, exploiting the lexical entries of our dictionaries.

Our future plans involve several tasks. First, we intend to make our dictionaries and morphological analyzers freely available for the speakers of Udmurt and Mansi and for anyone else interested in them. Second, we want to annotate our corpora with morphological and possibly syntactic information, which might serve as training data for statistical POS-taggers and syntactic parsers. Third, we also want to create online linguistic games that might help the process of language learning. We believe that the results achieved by the FinUgRevita project will contribute to the revitalization of Udmurt and Mansi and the tools to be developed will help these languages establish their existence in the digital space as well.

Acknowledgments

This work was supported in part by the Finnish Academy of Sciences and the Hungarian National Research Fund, within the framework of the project *Computational tools for the revitalization of endangered Finno-Ugric minority languages (FinUgRevita)*. Project number: OTKA FNN 107883; AKA 267097.

References

- [1] Naomi S. Baron. *Always on: Language in an online and mobile world*. Oxford University Press, Oxford, 2008.
- [2] András Kornai. Digital language death. *PLoS ONE*, 8(10):e77056, 2013.
- [3] István Kozmács. *Udmurt-magyar szótár*. Savaria University Press, 2002.
- [4] Е. И. Ромбандеева and Е. А. Кузакова. *Словарь мансийско-русский и русско-мансийский*. Просвещение, Ленинград, 1982.
- [5] Е. И. Ромбандеева. *Русско-мансийский словарь*. Миралл, Санкт-Петербург, 2005.
- [6] B. Munkácsi and B. Kálmán. *Wogulisches Wörterbuch*. Akadémiai Kiadó, Budapest, 1986.
- [7] А. Н. Баландин and М. П. Вахрушева. *Мансийско-русский словарь с лексическими параллелями из южно-мансийского (кондинского) диалекта*. Просвещение, Ленинград, 1958.

- [8] N. Thieberger and A. L. Berez. Linguistic data management. In N. Thieberger, editor, *The Oxford Handbook of Linguistic Fieldwork*, chapter 4, pages 90–118. Oxford University Press, Oxford, 2012.
- [9] Gábor Prószéky. Endangered uralic languages and language technologies. In *Proceedings of the Workshop on Language Technologies for Digital Humanities and Cultural Heritage*, pages 1–2, Hissar, Bulgaria, September 2011.
- [10] B. Kálmán. *Chrestomathia Vogulica*. Tankönyvkiadó, Budapest, 1963.
- [11] Béla Kálmán. *Wogulische Texte mit einem Glossar*. Akadémiai Kiadó, Budapest, 1976.
- [12] T. Riese. *Vogul*. Number 158 in Languages of the World/Materials. Lincom Europa, München - New Castle, 2001.
- [13] Е. И. Ромбандеева. *Мансийский (вогульский) язык*. Наука, Москва, 1973.
- [14] Е. И. Ромбандеева. *Графика, орфография и пунктуация мансийского языка*. Правительство Ханты-Мансийского Автономного Округа - Югры - Департамент Образования и Науки, Департамент по Вопросам Малочисленных Народов Севера, Обско-Угорский Институт Прикладных Исследований и Разработок, Ханты-Мансийск, 2006.